

<https://helda.helsinki.fi>

---

## Understanding the use of urban green spaces from user-generated geographic information

Heikinheimo, Vuokko Vilhelmiina

2020-09

---

Heikinheimo, V V, Tenkanen, H, Bergroth, C, Järvi, O, Hiippala, T & Toivonen, T 2020  
, ' Understanding the use of urban green spaces from user-generated geographic  
information ', Landscape and Urban Planning, vol. 201, no. Sept. 2020, 103845. <https://doi.org/10.1016/j.landurbplan.2020.103845>

---

<http://hdl.handle.net/10138/318851>

<https://doi.org/10.1016/j.landurbplan.2020.103845>

---

cc\_by

publishedVersion

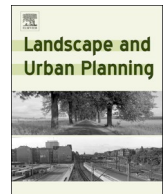
---

*Downloaded from Helda, University of Helsinki institutional repository.*

*This is an electronic reprint of the original article.*

*This reprint may differ from the original in pagination and typographic detail.*

*Please cite the original version.*



# Understanding the use of urban green spaces from user-generated geographic information



Vuokko Heikinheimo<sup>a,d,\*</sup>, Henrikki Tenkanen<sup>a,d</sup>, Claudia Bergroth<sup>a,c</sup>, Olle Järvi<sup>a,d</sup>,  
Tuomo Hiippala<sup>a,b</sup>, Tuuli Toivonen<sup>a,d,\*</sup>

<sup>a</sup> Digital Geography Lab, Department of Geosciences and Geography, University of Helsinki, Finland

<sup>b</sup> Department of Languages, University of Helsinki, Finland

<sup>c</sup> City of Helsinki, Finland

<sup>d</sup> Urbaria, Helsinki Institute of Sustainability Science, University of Helsinki, Helsinki, Finland

## ARTICLE INFO

### Keywords:

Urban green space  
Social media data  
Sports tracking data  
Mobile phone data  
PPGIS

## ABSTRACT

Parks and other green spaces are an important part of sustainable, healthy and socially equal urban environment. Urban planning and green space management benefit from information about green space use and values, but such data are often scarce and laborious to collect. Temporally dynamic geographic information generated by different mobile devices and social media platforms are a promising source of data for studying green spaces. User-generated data have, however, platform specific characteristics that limit their potential use. In this article, we compare the ability of different user-generated data sets to provide information on where, when and how people use and value urban green spaces. We compare four types of data: social media, sports tracking, mobile phone operator and public participation geographic information systems (PPGIS) data in a case study from Helsinki, Finland. Our results show that user-generated geographic information sources provide useful insights about being in, moving through and perceiving urban green spaces, as long as evident limitations and sample biases are acknowledged. Social media data highlight patterns of leisure time activities and allow further content analysis. Sports tracking data and mobile phone data capture green space use at different times of the day, including commuting through the parks. PPGIS studies allow asking specific questions from active participants, but might be limited in spatial and temporal extent. Combining information from multiple user-generated data sets complements traditional data sources and provides a more comprehensive understanding of green space use and preferences.

## 1. Introduction

Urban green spaces are crucial both for the well-being of people and for environmental sustainability in cities (Jenks & Jones, 2009; Lee & Maheswaran, 2011; MEA, 2005). Green spaces – the network of parks, forests and other green areas in an urban structure – provide opportunities for physical exercise and recreation, in addition to various other benefits to people (Gómez-Baggethun et al., 2013; Tyrväinen, Mäkinen, & Schipperijn, 2007). Access to and availability of green spaces are related to social and environmental justice (Ngom, Gosselin, & Blais, 2016; Wolch, Byrne, & Newell, 2014) and physical and mental health (Engemann et al., 2019; Tomita et al., 2017; Ward Thompson et al., 2012), even if the linkages are sometimes complex (Anguelovski, Cole, Connolly, & Triguero-Mas, 2018).

While the urban population of the world is growing rapidly (DESA/

UN-WUP, 2018), urban areas need to expand or densify in order to accommodate growing population, which is often carried out at the expense of green spaces (Haaland & van den Bosch, 2015; Zhou & Wang, 2011). Planning sustainable and socially equal cities requires knowledge about how people value and use urban green spaces (Burkhard, Kroll, Nedkov, & Müller, 2012). For example, understanding the importance of urban green spaces for recreation and social interaction is important when carrying out land use planning under urban development pressure (Haaland & van den Bosch, 2015).

Obtaining comprehensive data on the use of and values related to green spaces is challenging. Recreational use and preferences related to urban parks have been studied using questionnaire surveys (Tyrväinen et al., 2007), activity diaries (Mytton, Townsend, Rutter, & Foster, 2012), and more recently using GPS-based campaigns (Jankowska, Schipperijn, & Kerr, 2015) and map-based surveys based on public

\* Corresponding authors.

E-mail addresses: [vuokko.heikinheimo@helsinki.fi](mailto:vuokko.heikinheimo@helsinki.fi) (V. Heikinheimo), [tuuli.toivonen@helsinki.fi](mailto:tuuli.toivonen@helsinki.fi) (T. Toivonen).

<https://doi.org/10.1016/j.landurbplan.2020.103845>

Received 8 October 2019; Received in revised form 14 January 2020; Accepted 4 May 2020

0169-2046/ © 2020 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

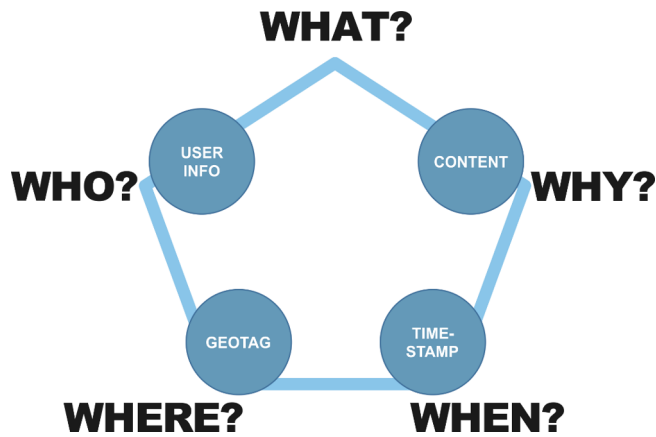


Fig. 1. User-generated geographic information provides insights about when and where people use green spaces, what they do there, why and who are using these spaces. Modified from Di Minin et al. (2015); Heikinheimo et al. (2017).

participation geographic information systems (PPGIS) (Brown & Kyttä, 2014; Brown, Schebella, & Weber, 2014; Laatikainen, Tenkanen, Kyttä, & Toivonen, 2015). These approaches provide in-depth understanding about green space use, but are often limited in duration and frequency, because data collection that involves active participation can be time-consuming for everyone involved.

On the other hand, we witness new digital data sources emerging from different types of crowdsourcing initiatives, citizen science projects and big data feeds. Crowdsourcing and citizen science projects gather information and insights from active contributors via online tools (See et al., 2016), whereas big data refer to overwhelming amounts of diverse information produced constantly by and about people through online networks and different digital sensors (boyd et al., 2012; Kitchin, 2014). For instance, GPS-enabled mobile devices and online platforms hosting geolocated user-generated content provide detailed information about the whereabouts and

activities of people over time in large quantities. Different elements of user-generated data sets, such as geotags, timestamps, content and user information, provide new possibilities for studying the use of green spaces from different perspectives (Di Minin, Tenkanen, & Toivonen, 2015) (Fig. 1). These data provide new opportunities for understanding how cities function in space and time (Batty, 2013) and how human-nature interactions occur in different environments (Toivonen et al., 2019). There are certainly theoretical, technical and ethical challenges for applying these new data sources, however, even imperfect measures – when used with caution and critical thinking – are better than no consideration of the value of nature in decision making (Daily, Postel, Bawa, & Kaufman, 1997).

The aim of this paper is to discuss what kinds of information about urban green space use and values can be extracted from different types of user-generated geographic information and applied to urban green space planning. We use the concept of user-generated geographic information to capture different kinds of novel digital data sources that involve the interaction of people and location-based technologies. We compare four different types of data for analysing green space use and preferences: social media data, sports tracking data, mobile phone operator data and PPGIS data.

Stemming from the framework proposed by Di Minin et al. (2015) (Fig. 1), we study Helsinki, Finland, as our case area and explore: 1) where the spatial hotspots of green space use are, 2) when people use green spaces, 3) what activities are present in green spaces and 4) who are using green spaces based on available sample data sets. Finally, we compare the different data sets as a source of knowledge about urban green space usage and we discuss the potential and challenges of using user-generated data to inform sustainable spatial planning and green space management in cities.

## 2. Characteristics of user-generated geographic information

The data sets used in this study represent various kinds of user-generated geographic information ranging from passively contributed to actively contributed data (See et al., 2016). Social media data and mobile phone data are originally generated for other purposes than

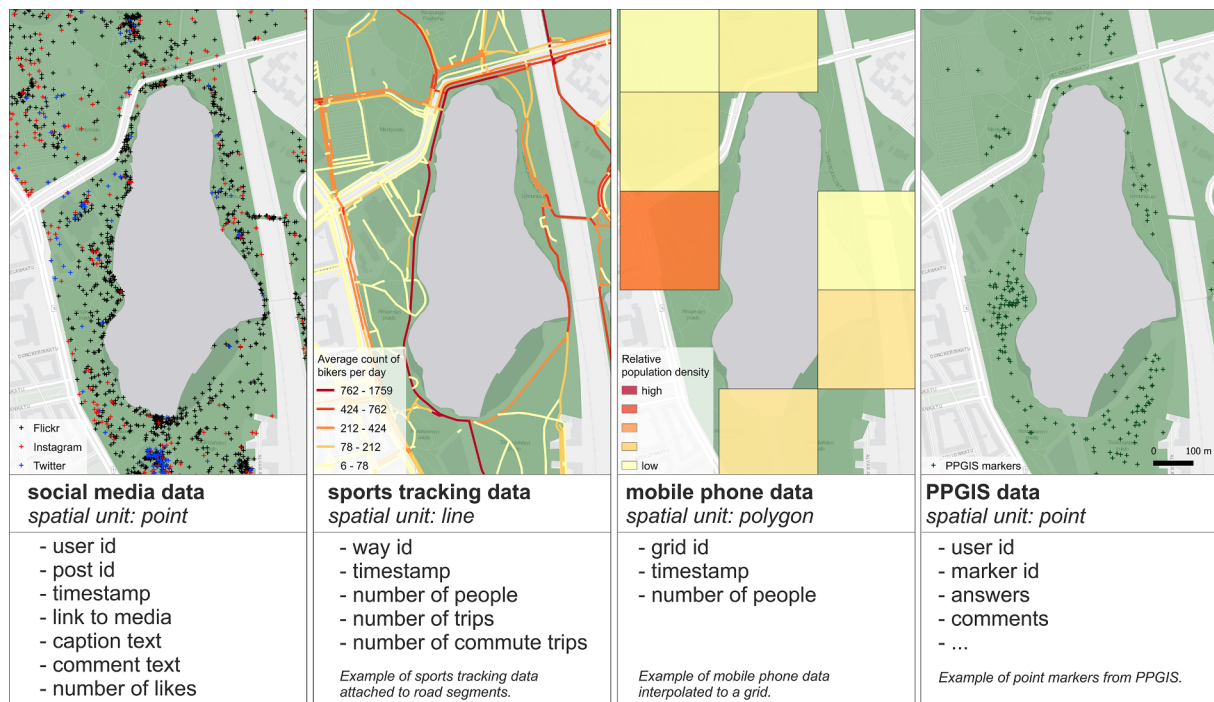


Fig. 2. Examples of a) social media data, b) sports tracking data, c) mobile phone data and d) PPGIS data from green spaces in Helsinki, Finland. Social media data and PPGIS data represent the original user-generated points, and the spatial accuracy of these points may vary from exact to coarse. Sports tracking data and mobile phone data are in an aggregated format. In addition to varying spatial units, the attributes of the data sources also differ. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

research or volunteered mapping efforts, and thus they fall into the category of passively or indirectly contributed data. In contrast, PPGIS data are generated by users who actively participate in a map-based survey, and the users are aware that the data will be used for planning or research. Sports tracking data, and other similar GPS-based activity tracking data often fall into the category of passively contributed data if acquired from the data provider in large quantities, but can also be gathered as smaller samples from active study participants.

While different types of user-generated data sets describe people's whereabouts, activities or preferences to some extent, they differ in technical details and thematic content (Fig. 2). These properties affect the way how empirical analysis can be done and what conclusions made based on each data set. Passively contributed data are often more voluminous, but might lack necessary background information and other details. Actively contributed data often allows the collection of relevant details, but might be limited in duration and extent (Levin, Lechner, & Brown, 2017).

Spatial accuracy is a key aspect when analysing user-generated geographic information. Social media data might be attached to exact coordinates, or more coarse points-of-interests based on place names (Hochmair, Juhász, & Cvetojevic, 2018; Toivonen et al., 2019). Sports tracking data is originally captured as exact GPS points, but often delivered in aggregate format. Allocating mobile phones to locations depend on the antenna network (Ahas, Silm, Järv, Saluveer, Tiru, 2010; Järv, Tenkanen, & Toivonen, 2017). PPGIS data often represents markers that users have placed on a map in a web browser, and the preciseness can depend on the zoom-level and local knowledge of the user (Brown, 2012).

### 2.1. Social media data

Social media refers to web-based services that allow people to create and share content in online communities (McCay-Peet & Quan-Haase, 2017). Flickr and Twitter are the most used social media data sources in environmental studies (Ghermandi & Sinclair, 2019), mostly because of the availability of publicly available data via the platforms' application programming interfaces (APIs). Panoramio, Instagram, Facebook and Foursquare have also been popular data sources for studying human activities and presence in nature. While most environmental studies have used data from a single platform (Ghermandi & Sinclair, 2019), combinations of data sources may produce the best understanding of green space visitors (Tenkanen et al., 2017).

There has been a recent rapid increase in using social media for studying people's activities and preferences in geographical research. Location-based social media data are used widely in urban research ranging from urban form and structure to activity practices of people (Crooks et al., 2015; Huang & Wong, 2016; Ilieva & McPhearson, 2018; Shelton, Poorthuis, & Zook, 2015; Steiger, Westerholt, Resch, & Zipf, 2015) and increasingly in environmental studies (Ghermandi & Sinclair, 2019; Toivonen et al., 2019). Studies have found that social media usage rates reflect observed visitation rates in popular nature destinations such as national parks (Tenkanen et al., 2017; Wood, Guerry, Silver, & Lacayo, 2013) and urban parks in large cities (Donahue et al., 2018; Hamstead et al., 2018).

### 2.2. Sports tracking data

Sports tracking applications allow people to trace their physical activities using satellite navigation and mobile devices. Sports tracking applications, such as Strava, Sports Tracker and MapMyFitness track the user's activities by combining information from different sensors, such as GPS, accelerometer and heart rate monitor, and sometimes also manually entered information about the activity (Lendák, 2016).

Sports tracking data may contain information about recreational activities, such as biking and walking, and utilitarian activities, such as commuting to and from work (Oksanen, Bergman, Sainio, & Westerholm, 2015). Previous studies have used GPS-based activity data to study the

relationship of physical activity and urban green infrastructure (Vich, Marquet, & Miralles-Guasch, 2019), as well as on- and off-trail use in protected areas (Norman & Pickering, 2017) and urban parks (Korpilo, Virtanen, & Lehvävirta, 2017). Previous studies found similar patterns in sports tracking data and manual cycling counts (Jestico, Nelson, & Winters, 2016; Oksanen et al., 2015), but sports tracking data is also known to be biased towards men and active athletes.

### 2.3. Mobile phone data

Studies using mobile phone data are often based on passive mobile positioning data (as opposed to active tracing of mobile phones), which refers to information that mobile phone operators automatically store in log files (Ahas, Silm, et al., 2010). Mobile phone data contain information about the location of mobile devices in the mobile phone operator's network – to which network antenna a mobile phone is connected to (Ahas, Silm, et al., 2010). Mobile phone data may contain information about calls, text messages, data usage and network connection attempts. As the network antennae are situated with varying densities, the spatial accuracy of mobile phone data can vary. However, latter can be enhanced through spatial interpolation (Järv et al., 2017).

Mobile phone data have been used as a proxy of population distribution and movements of people to reveal population and urban dynamics over time (Ahas, Silm, et al., 2010; Becker et al., 2011; Järv, Tenkanen, Salonen, Ahas, & Toivonen, 2018). These data have been further applied to transportation (Iqbal, Choudhury, Wang, & González, 2014), tourism (Ahas, Aasa, Roose, Mark, & Silm, 2008; Fisher et al., 2019), social inequality (Järv, Muurisepp, Ahas, Derudder, & Witlox, 2015) and health research (Dewulf et al., 2016) but also to studying urban green spaces (Guo et al., 2019).

### 2.4. PPGIS data

Public participation geographic information systems (PPGIS) and participatory geographic information systems (PGIS) refer to approaches that combine participatory methods with geographic information technologies to collect insights from the general public, often in the context of a planning process (Brown & Kyttä, 2014). PPGIS surveys often allow users to mark places, routes and areas on a (web-based) map when answering a set of questions. Responses may be gathered from household sampling groups or volunteer participants (Brown, 2017).

PPGIS studies have focused, for example, on values for green spaces (Ives et al., 2017; Pietrzyk-Kaszyńska, Czepkiewicz, & Kronenberg, 2017), accessibility of recreation by the water (Laatikainen et al., 2015), physical activities and benefits from parks (Brown et al., 2014), perceived environmental quality (Kyttä, Broberg, Tzoulas, & Snabb, 2013; Wang, Kotze, Vierikko, & Niemelä, 2019) and urban happiness (Kyttä, Broberg, Haybatollahi, & Schmidt-Thomé, 2016).

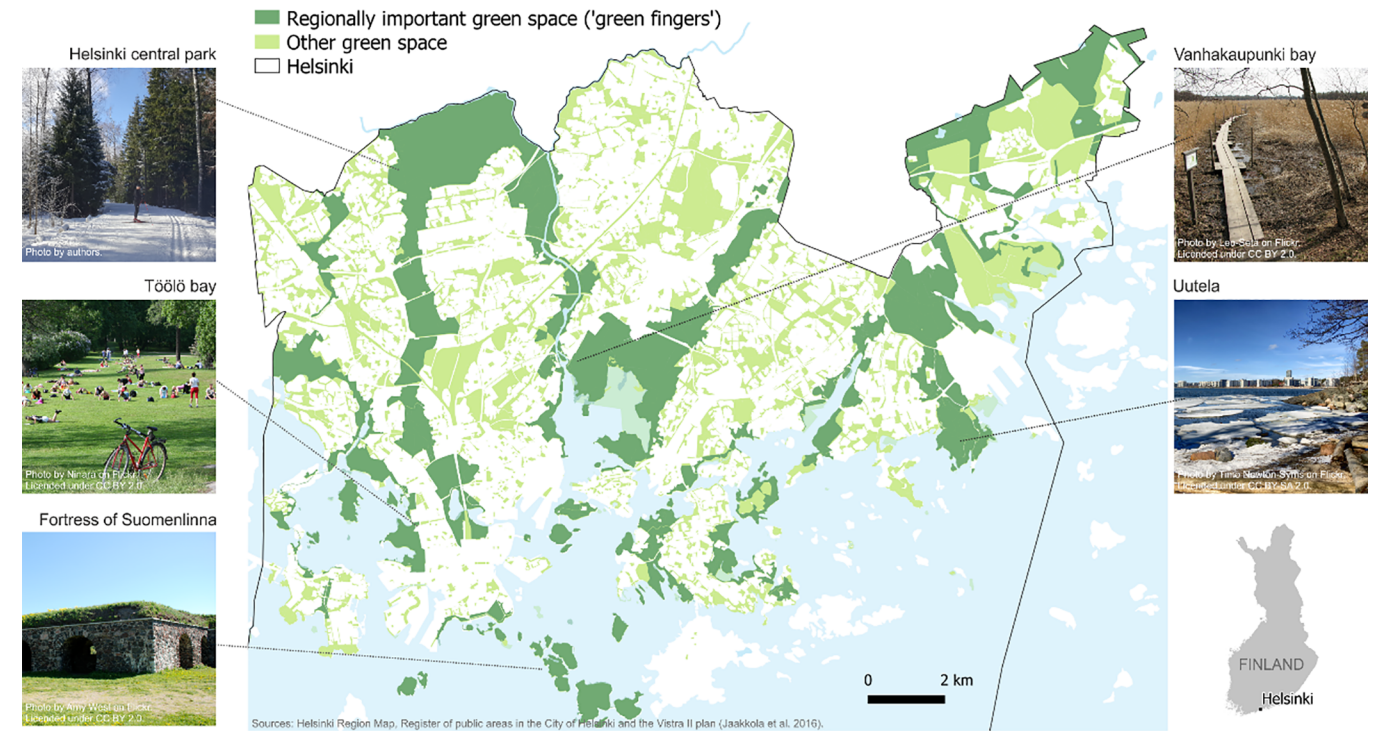
## 3. Material & methods

### 3.1. Green spaces in Helsinki

Our sample data cover Helsinki, the capital of Finland. Helsinki is a medium-sized city with a population of 648 042 in 2018 (Statistics Finland, 2019). In addition to the inhabitants, the city is visited by tourists, who also contribute to the pool of user-generated data. In 2017, the city recorded 4.2 million overnight stays in Helsinki, 46% by domestic and 54% by foreign travellers (City of Helsinki, 2018).

Densification of the city structure is one of the major goals in the general plan of Helsinki, and strategic planning of the green space network aims to ensure a multifunctional and connected green infrastructure into the future (Jaakkola, Böhling, Nicklén, & Lämsä, 2016). From a regional perspective, the core of the green space network in Helsinki is the so-called 'green fingers': six sections of green infrastructure that extend from the coast towards the north (Fig. 3).





**Fig. 3.** Public green spaces in Helsinki, as delineated in this study (City of Helsinki, 2019; Jaakkola et al., 2016). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

In this study, we used green space polygons from the register of public areas in Helsinki (City of Helsinki, 2019), and we further identified regionally important green spaces using a polygon layer to delineate the green fingers (Jaakkola et al., 2016). We used 250 m x 250 m statistical grid squares that have their centroid inside the green finger polygons in the grid-based comparisons.

3.2. Sample data sets

We gathered available user-generated data from various sources and subset all data sets to the extent of Helsinki (Table 1). More details about the data sets and data processing are available in the Supplementary materials. Additional documentation and Python scripts used for producing results in this paper are available online at: <https://github.com/DigitalGeographyLab/some-urbangreens>.

**Social media data** (posts) were collected from three different platforms: Flickr, Instagram and Twitter. We accessed publicly available geo-tagged social media data from Flickr API ([www.flickr.com/services/api/](http://www.flickr.com/services/api/)), Instagram API ([www.instagram.com/developer/](http://www.instagram.com/developer/)), and Twitter API ([developer.twitter.com](http://developer.twitter.com)). We subset the data for one calendar year – 2015 for Flickr and Instagram and 2017 for Twitter – according to data availability, and identified points that spatially intersect with the green spaces. We also calculated the number of social media users in each 250 m x 250

m analysis grid square. See further details about social media data collection and pre-processing in the Supplementary materials (S1–S3).

**Sports tracking data** from the Strava platform were acquired as an aggregated data set, which contains information about the number of athletes and trips and commuting trips per minute in each road segment (Strava Metro 2016; Tarnanen, Salonen, Willberg, & Toivonen, 2017). We identified road segments that intersected with the green spaces using a spatial overlay. We joined information about number of athletes and trips to the 250 m x 250 m analysis grid by selecting the maximum number of athletes from intersecting road segments for each grid square. See further details about the Strava data user base in the Supplementary material (S4).

Passively collected **mobile phone data** from a two-and-a-half-month period (28.10.2017–9.1.2018) was acquired from a major Finnish mobile network operator company. The mobile phone data used in this study are based on the number of hourly data use attempts (e.g. browsing the internet on a mobile device or email synchronization) made by the users in the mobile network (Bergroth, 2019). The data set was anonymized by the mobile network operator, and we aggregated data from regular weekdays (Monday to Thursday) into 250 m x 250 m statistical grid squares on an hourly interval prior to further analysis (Bergroth, 2019). Each grid square in the final data set contains information about the relative share of estimated population across the

**Table 1**  
Details of the user-generated data sets covering the City of Helsinki.

Data set name	Type	Source	Time period (dd/mm/yyyy)	Records	Users
Flickr	Social media	Flickr API	1/1/2015–31/12/2015	29 287	902
Instagram	Social media	Instagram API	1/1/2015–31/12/2015	602 466	113 754
Twitter	Social media	Twitter API	1/1/2017–31/12/2017	31 359	5386
PPGIS2050	PPGIS	City of Helsinki: Helsinki 2050 survey	4/11/2013–9/12/2013	6307/28250*	1665/2588*
PPGISPark	PPGIS	City of Helsinki: Questionnaire about Helsinki's national city park	25/10/2017–17/12/2017	10 939	1385
Strava	Sports tracking application	Strava METRO data set	1/1/2015–31/12/2015	161 946	4044
Mobile phone data	Mobile phone operator data	Elisa Oyj; Bergroth, 2019	28/10/2017–9/1/2018	**	**

\* Green-space-related answers/all answers within Helsinki.

\*\* We had no information about the absolute value of records or users, only proportional values across the whole region.

**Table 2**

Proportion of each data set in public green spaces. Percentages indicate the share of data in green spaces out of data in the whole city, absolute numbers in parenthesis. For Strava data and mobile phone data, exact number of users and records were only available in the metadata for the whole data set.

Name	Records in green spaces	Users in green spaces	Average number of records per user in green spaces
Flickr	22% (6302/29 287)	48% (432/902)	14.6
Instagram	15% (90 211/602 466)	28% (32 309/113 754)	2.8
Twitter	6% (1939/31 359)	16% (882/5386)	2.2
PPGISPark	18% (5164/28 250)	59% (1518/2588)	3.4
PPGIS2050	67% (7320/10 939)	89% (1238/1385)	5.9
Strava	8%	7%	–
Mobile phone data	–	3%	–

whole region. For green spaces, we considered mobile phone data from grid squares having their centroid inside the green finger polygons. In spatial and temporal comparisons, we used the information from 4 p.m. to 5 p.m., because this time interval received the most activity across green spaces in the mobile phone data set. For more details about the data processing, see [Supplementary material \(S5\)](#) and [Bergroth \(2019\)](#).

**PPGIS data** was acquired from two surveys conducted by the local municipality: the Helsinki 2050 survey from 2013 and the National Urban Park Survey from 2017. The Helsinki 2050 survey was conducted in order to support the development of an upcoming general plan for the City of Helsinki ([Kahila-Tani, Broberg, Kyttä, & Tyger, 2016](#)). The survey contained 16 question in total, out of which we considered two questions related to green spaces. The National Urban Park Survey was designed to support the planning of a national urban park in Helsinki – a project that aims to secure cultural and landscape values in cities. We considered all 12 questions that allowed users to add point markers on the map. For both data sets, we only considered PPGIS point markers located in Helsinki, and further only those markers intersecting the green spaces ([Table 1](#)). We also calculated the number of PPGIS users in each 250 m × 250 m analysis grid square. Both PPGIS data sets are openly available online in the Helsinki Region Infoshare open data portal ([https://hri.fi/en\\_gb/](https://hri.fi/en_gb/)). See further details about the PPGIS data sets in the [Supplementary material \(S6\)](#).

### 3.3. Spatial and temporal analysis

We first inspected what proportion of each data set was produced in green spaces, in comparison with data across the whole city. In further spatial comparisons we used the 250 m × 250 m grid squares that have their centroid inside the green finger polygons in order to make the other data sets comparable with the mobile phone data.

We identified hotspots as the top quintile (20% of grid squares) of each data set, based on the number of users in the green space grid squares. We compared the spatial hotspots using the Jaccard index, following the approach in ([Lehtomäki, Tuominen, Toivonen, & Leinonen, 2015](#)). The Jaccard index measures the similarity between two sets and is calculated as the intersection of two sets divided by their union. Value 1 indicates complete overlap between the two sets, and value 0 means the two sets do not overlap.

We compared the temporal patterns of social media data, sports tracking data and mobile phone data from green spaces using temporal plots that show the relative share of user activity at each time unit. We aggregated the data per hour of the day, weekday and month, when possible, to allow for visual comparison of the temporal patterns.

### 3.4. Content analysis of social media data

We conducted manual content classification for a subset of the social media images from Flickr and Instagram. We selected a sample of social media posts for the content classification based on a spatial intersection between the social media points and a coarse-resolution polygon of the regionally important green spaces ('green fingers') retrieved from a regional land use plan from the Helsinki-Uusimaa Regional Council. See further details about the data pre-processing in the [Supplementary material \(S2\)](#). In total, we manually classified the content of 15 312 Instagram photos and 1843 Flickr photos.

The objectives of the content analysis were to assess data quality by determining how many of the photos located in green spaces actually contained information about green spaces, and to understand what activities from green spaces are portrayed on social media. Advertisements and photos taken inside people's homes were determined as not relevant content. Content of relevant images from green spaces were further labelled if they contained activities (such as jogging or biking) or a landscape (a picture with a visible horizon and/or a wide view of the landscape). One picture could be allocated to several categories.

### 3.5. Language identification

Strava data and PPGIS contained some metadata about the data producers (see [Supplementary materials S4 and S6](#)), while social media data and mobile phone data lacked any additional information about who produced the data. We chose to demonstrate how language identification could improve our understanding of the data producers and green space visitors.

We limited language identification to the most extensive social media data set available, namely Instagram. Instagram posts have relatively long caption texts, which improves automatic language identification ([Baldwin & Lui, 2010](#)).

We identified the language of all posts by first pre-processing the captions and segmenting them into sentences, as proposed in ([Hiippala, Hausmann, Tenkanen, & Toivonen, 2019](#)). We then used a pre-trained model capable of identifying 176 languages ([Bojanowski, Grave, Joulin, & Mikolov, 2016](#)) to identify the language of each sentence. We excluded both short sentences (< 7 characters) and those identified with low confidence (< 0.5). We then summarized the language information for each user. If the user had used more than one language, we first excluded English from the list of languages identified, as social media users often post in English in addition to their first language ([Hiippala et al., 2019](#)). If more than one language remained after removing English, we assumed that the most frequently used language corresponded to the user's primary or first language. For the final analysis, we only considered languages with more than 10 users in our data set. Finally, we summarized the number of different languages in the 250 m × 250 m analysis grid squares across green spaces.

## 4. Results

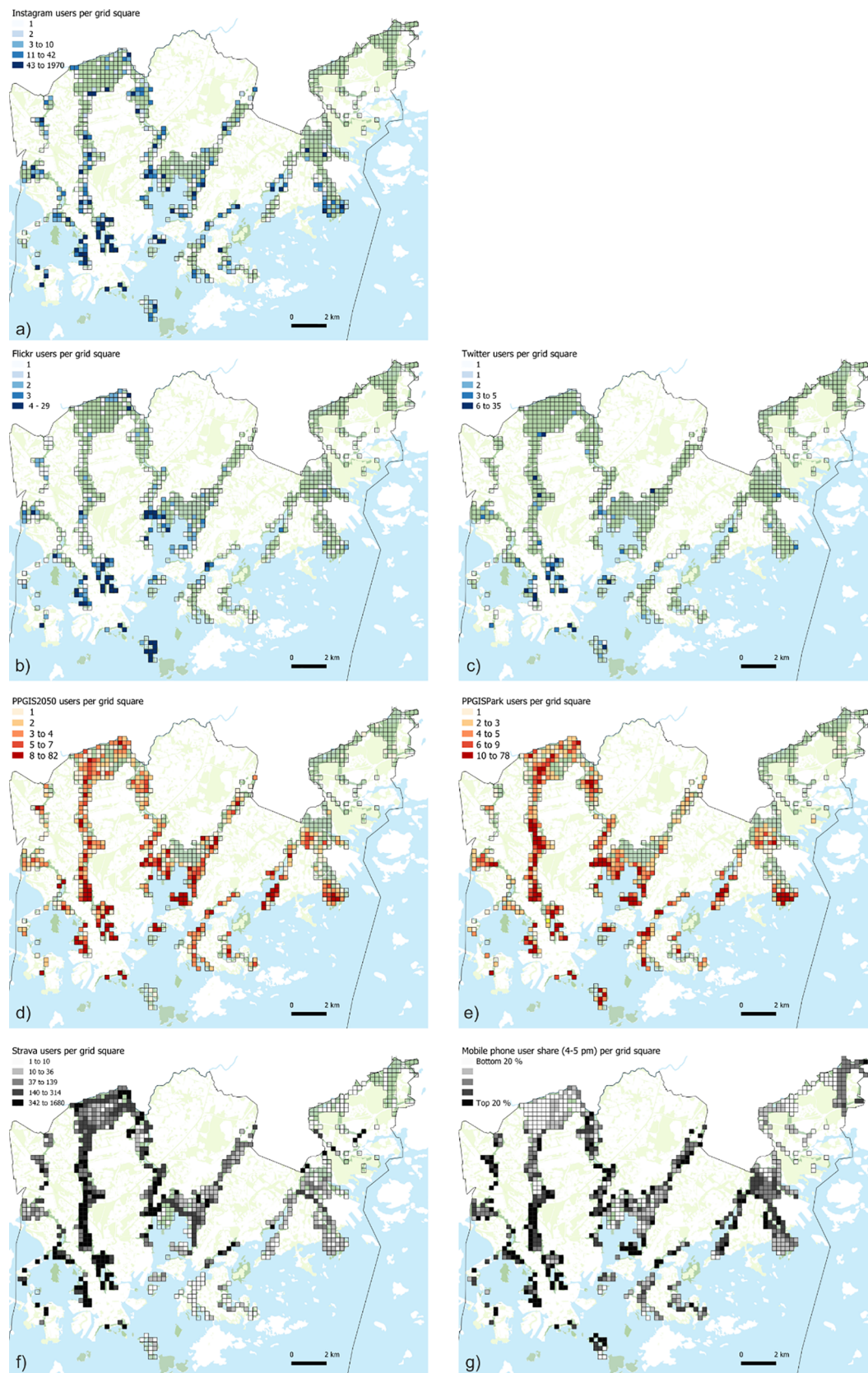
### 4.1. Spatial patterns of green space use

*Where are the spatial hotspots of green space use?* The volume and proportion of information from green spaces varied between the data sets ([Table 2](#)). In absolute numbers, Instagram was the most extensive data set across the city and from green spaces, while Flickr had clearly the highest ratio of records (posts) per user inside and outside green spaces ([Table 2](#)). PPGIS data sets had the highest proportion of posts located in green spaces, which is logical because the PPGIS questionnaires included direct questions about nature and recreation.

Our sample data sets for sports tracking data and mobile phone data did not contain spatially explicit information about the absolute number of individual users in specific areas. Based on the aggregated data sets across the city, we calculated that the proportion of users in green spaces was 7% in the Strava data and 3% for mobile phone data.

The spatial pattern of users in green spaces was rather different



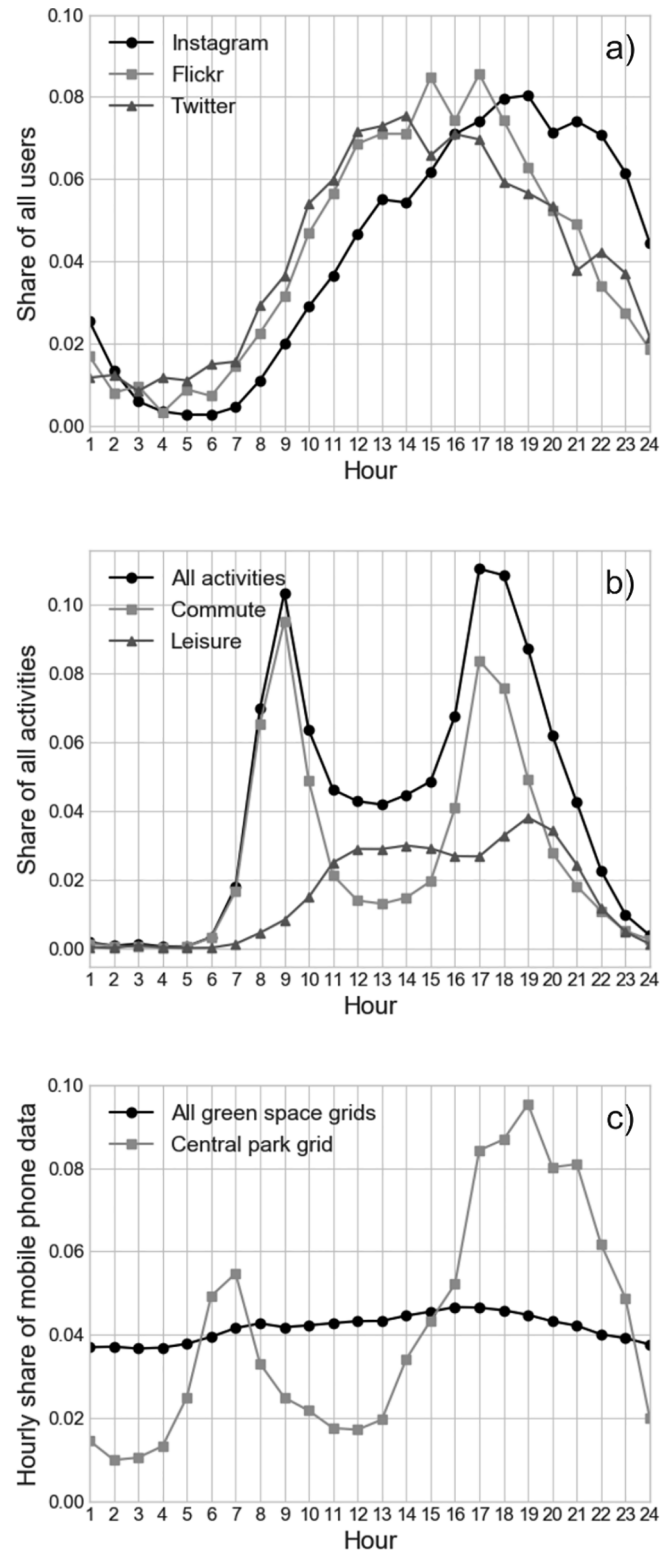


**Fig. 4.** Spatial distribution of users in the different data sources represented as quantiles in the 250 m × 250 m grids: a) Instagram, b) Flickr, c) Twitter, d) PPGIS 2050, e) PPGISPark, f) Strava, g) Mobile phone data. The top quantile (20% of data) of each data set were considered as hotpots. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

**Table 3**  
Jaccard similarity coefficient between hotspots of the different user-generated data sets in green spaces. Value 1 would indicate complete overlap and value 0 no overlap.

	PPGIS2050	PPGISPark	Instagram	Flickr	Twitter	Mobile Phone	Strava
PPGIS2050	1.00						
PPGISPark	0.37	1.00					
Instagram	0.20	0.20	1.00				
Flickr	0.16	0.23	0.30	1.00			
Twitter	0.14	0.11	0.25	0.20	1.00		
Mobile Phone	0.15	0.13	0.14	0.12	0.06	1.00	
Strava	0.16	0.18	0.18	0.14	0.07	0.23	1.00

among the data sources (Fig. 4). The jaccard index between data hotspots was below 0.40 among all data sets (Fig. 4; Table 3), also when testing with different threshold values for the hotspots (see Supplementary materials, tables S1 and S2). The most similar hotspot patterns were between the two PPGIS data sets ( $J = 0.37$ ). Furthermore, hotspots of the three different social media data sets were the most similar among each other ( $J = 0.30$  between Instagram and



**Fig. 5.** 24-hour patterns in a) social media data, b) sports tracking data and c) mobile phone data from green spaces.



Flickr,  $J = 0.25$  between Instagram and Twitter and  $J = 0.20$  between Flickr and Twitter). Between different types of data, the most similar hotspot distributions were between PPGISPark and Flickr ( $J = 0.23$ ), mobile phone data and Strava sports tracking application data ( $J = 0.23$ ) and PPGIS2050 and Instagram ( $J = 0.20$ ). Overlap among the hotspots from different data sets are visualized in the [Supplementary materials \(Fig. S1 and S2\)](#).

#### 4.2. Temporal patterns of green space use

When are people using green spaces? Temporal patterns based on social media data, sports tracking data and mobile phone operator data reveal patterns of leisure time park use and commuting.

Aggregated 24-hour patterns of social media data extracted from green spaces show that most of the users share content in the afternoon and evening (Fig. 5a). Contrastingly, sports tracking data from green spaces show a clear diurnal activity pattern with peaks in the morning and afternoon (Fig. 5b). When looking only at the non-commuter trips

(‘leisure trips’) in the Strava data, there is an activity peak in the late afternoon/early evening (Fig. 5b), similar to the pattern in social media data. Diurnal patterns are clear in the mobile phone data when inspecting individual grid squares in popular recreational and commuting areas, such as Central Park in Helsinki (Fig. 5c). When looking at all green space grid squares together, the share of users at each hour averages at around 4% (Fig. 5c).

Weekday observations were possible to carry out only using social media data and sports tracking data. Data grouped by weekdays show increased park use activity during the weekends both in social media data (Fig. 6a) and leisure trip data from Strava (Fig. 6b).

#### 4.3. Activities in green spaces

What are people doing in green spaces? The first step of our social media content analysis identified how much of the data subset from green spaces contained relevant information. For Flickr and Instagram, the majority of data in green spaces was classified as relevant (Flickr:

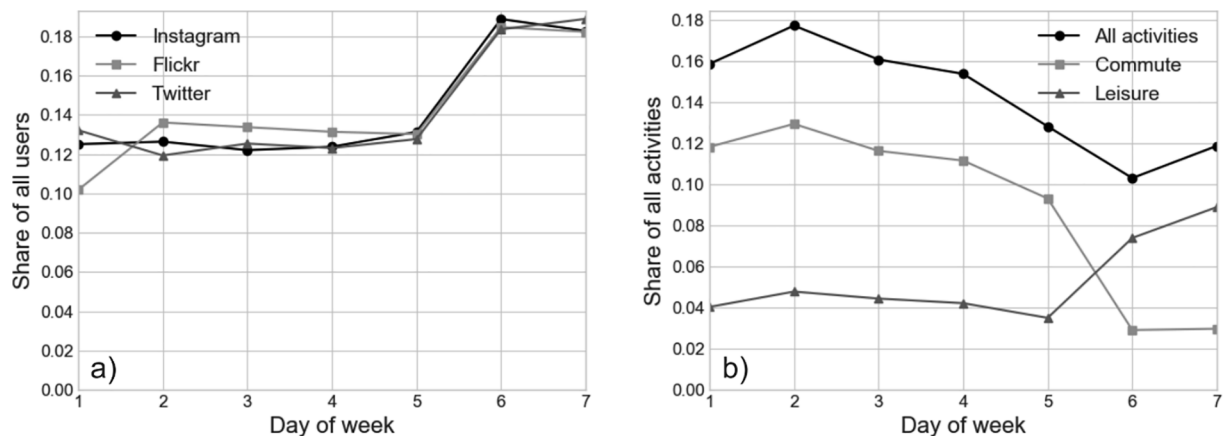


Fig. 6. Weekly temporal patterns in a) social media data (Instagram, Flickr, Twitter) and b) sports tracking data (Strava) from green spaces.

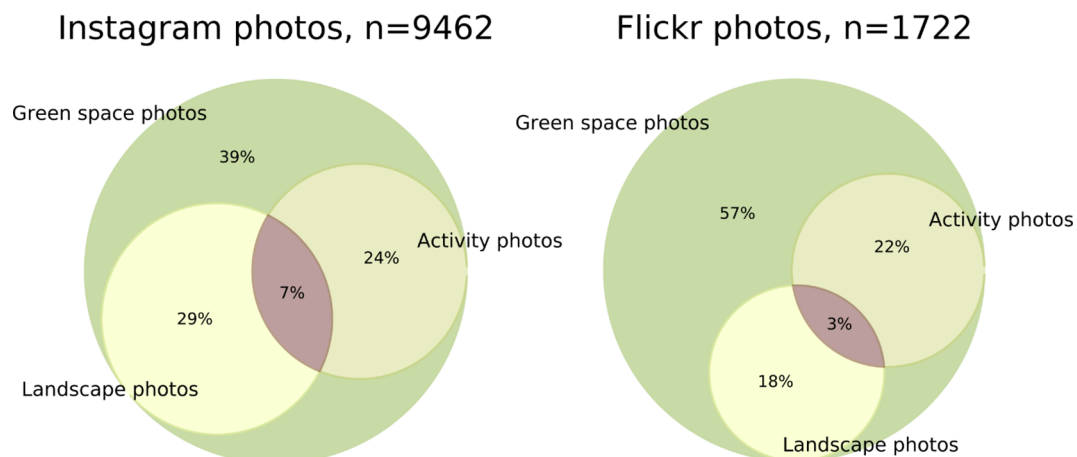
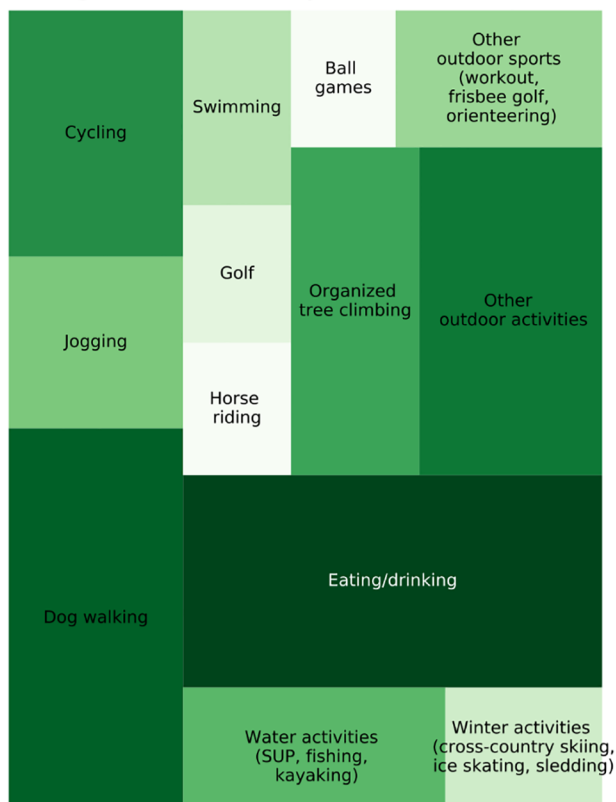
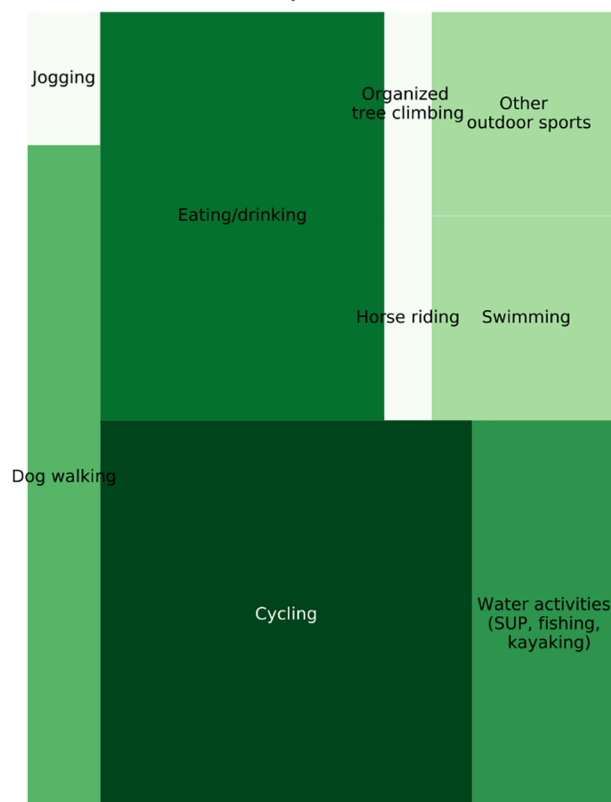


Fig. 7. An area-weighted Venn diagram showing the proportion and overlap of landscape photos and activity photos among all classified green space photos. ‘Green space photos’ not falling into the categories of ‘Activities’ or ‘Landscape photos’ include close-up photos of plants, pictures of buildings and photos of people posing for the camera in green spaces. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

## Instagram activities per number of users



## Flickr activities per number of users



**Fig. 8.** Treemap visualization of the number of Instagram users ( $n = 2086$ ) and Flickr users ( $n = 38$ ) per activity category. The size of the block corresponds to the relative proportion of the number of users that posted about each activity. Darkest color with white text represents the most popular activity per platform. Exact numbers of users and photos per activity category can be found in the [Supplementary materials Table S3](#).

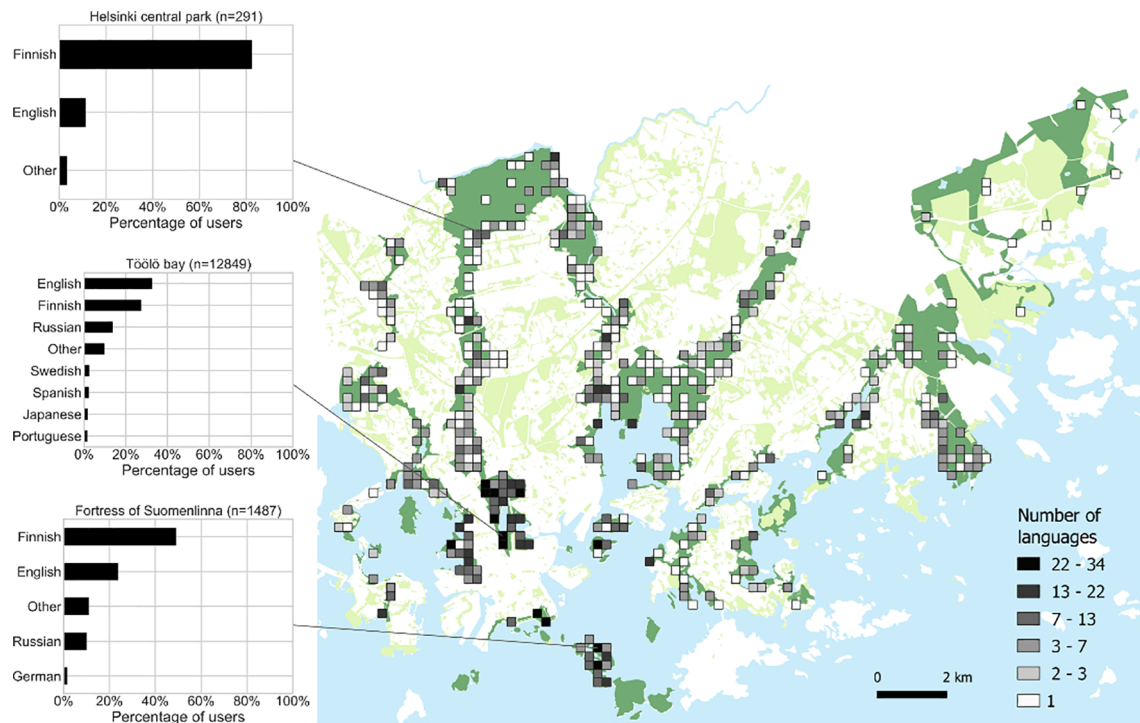
93%, Instagram: 83%), while the rest were classified as not relevant (Flickr: 5%, Instagram: 15%) or not available at the time of data analysis (Flickr: 1%, Instagram: 2%). We found that most of the geotagged Twitter data from green spaces (74%) had been originally posted on Instagram (based on the media links, see [Supplementary material S3](#) for details), and we did not conduct further content analysis on Twitter data.

Content analysis of images from regionally important green spaces fell into the main categories of landscape photos and activity photos rather similarly among Instagram and Flickr ([Fig. 7](#)). Flickr data contained a higher proportion of miscellaneous photos from green spaces including, for example, close-up images of plants. Mostly the same physical activities, such as cycling, jogging and dog walking, appeared in both Flickr and Instagram, but in different proportions ([Fig. 8](#)). Eating and drinking (for example, having a picnic) was also popular in both data sets. The activities observed from the Flickr data rely heavily on individual users as there were only 38 users who had shared photos of activities in the classified data set. For example, one user had shared 196 photos of jogging from a single event. Exact number of users and photos per activity category in the classified sample data can be found in the [Supplementary material \(Table S3\)](#). All in all, the Instagram data set was more diverse in terms of the activities present in the photo content.

#### 4.4. Languages in green spaces

*Who are visiting urban green spaces, and who are producing the data?* Language could be identified for 94% of users who had geotagged social media content to green spaces. Remaining users had posted content whose language could not be reliably identified (e.g. short sentences or low confidence) or did not include any linguistic content (e.g. captions consisting of emojis only). In total, we identified 37 languages used by at least 10 unique users in green spaces ([Appendix E](#)). The most frequent languages were Finnish and English, followed by Russian and Swedish. Less frequently used languages included Japanese, German, Italian, Spanish, Estonian and French. A full list of languages and their users is available in the [Supplementary materials \(Table S4\)](#).

When comparing the spatial pattern of linguistic richness ([Fig. 9](#)) with the spatial hotspots in general ([Fig. 4](#)), green spaces with more users and touristic attractions are clearly richer in terms of languages used. Linguistic richness decreases in green spaces further away from the city centre. Finnish was the most used language in most green spaces, except for locations that also serve as touristic destinations, such as Töölö Bay.



**Fig. 9.** The map shows the number of used languages across green spaces of Helsinki, based on Instagram data from 2015. Bar graphs show the proportion of different languages in selected grid squares from different kinds of green spaces. Languages with less than 1.5% share per grid square are grouped as “other” in the bar graphs. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

## 5. Discussion

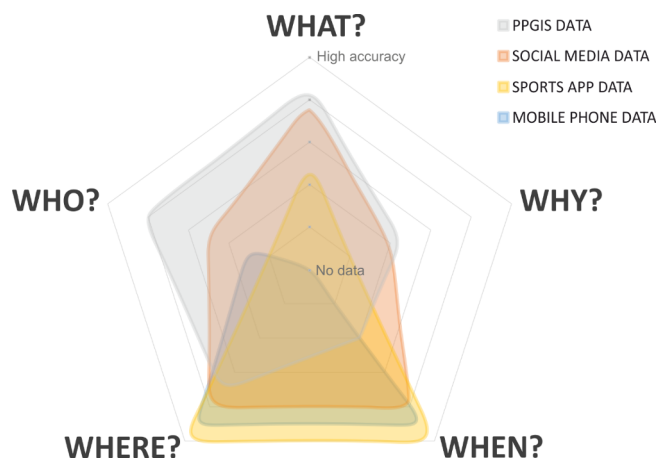
### 5.1. Similarities and differences among the user-generated data sources

Our results indicate that both active and passive user-generated geographic information have the potential to provide new insights about green space use and people’s preferences towards green spaces. Fig. 10 aims to summarize the commonalities and differences among social media data, sports tracking data, mobile phone data and PPGIS

from the perspective of understanding *where*, *when*, *why* and *in what way* people use green spaces and *who* these people are.

The radar chart describes the relative applicability of each data type for answering the different questions. We positioned each data source on the radar chart based on the sample data sets used in this study, and our evaluation of their applicability more broadly. For example, social media data, sports tracking data and mobile phone data provide rather high-quality data regarding the spatial and temporal aspects of green space use, but the amount of detail about activities in green spaces varies between data sources. Fitness for purpose of PPGIS data is very context specific as their information content depend strongly on the study design. PPGIS data often allows high quality data for understanding *who*, *what* and *why*, moderate accuracy about *where*, but little information about *when* in comparison to the other data types. Furthermore, the data sets can be used in combination with each other in order to gain a more comprehensive understanding about different questions and to potentially mitigate some of the biases in these data sets. Table 4 draws together main aspects of each data source that should be considered when using these data.

Research and decision making can gain different perspectives about green space use from these data. Social media data depict particularly the patterns of leisure time green space use, *being* in the parks (What are people posting about, when and where?). Sports tracking data and mobile phone data can also capture the daily rhythm of *moving through* the parks, including every-day commuting (Where and when are people moving in green spaces?). The PPGIS survey data used in this study focused on green space values, *perceiving* the parks, and not on actual green space use (Who are valuing specific green spaces and why?). PPGIS data differs from the other more passively produced data sources in general, as the surveys can be designed according to information needs of planners and researchers. When using these data sources for research and planning, there are several questions related to data access, information content and biases to consider (see more details in Table 4).



**Fig. 10.** An illustration of the potential applicability of different user-generated data sets for answering the questions where, when, what, why and who as interpreted by the authors. The further towards the corners of the radar chart, the higher the quality of the data for answering the different questions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

**Table 4**  
Comparison of different user-generated data sources for studying the use of urban green spaces.

Data type	Description	Accessing the data	Spatial coverage	Spatial accuracy	Temporal information	Activities and preferences	User information	Example questions	Known biases
Social media data	Data from social networking sites that allow users to share (georeferenced) content onlineFor example: Flickr, Instagram, Twitter, Facebook, Weibo	<ul style="list-style-type: none"> <li>- Browsing the data online</li> <li>- Programmatic access through APIs</li> <li>- Buying a data set</li> <li>- Access varies between platforms</li> </ul>	Potentially global extent depending on platform popularity.	Accuracy varies according to platform as user input as well as user's device: <ul style="list-style-type: none"> <li>- GPS accuracy points</li> <li>- Point of interest (POI)</li> <li>- Place names in text</li> </ul>	Timestamp reflects the time of uploading the content, not necessarily the actual time of the event.	Interpretation from textual and visual content.	Username, home location, use history, social network	<ul style="list-style-type: none"> <li>- WHERE and WHEN do tourists visit green spaces in Helsinki?</li> <li>- WHAT new activities have emerged?</li> </ul>	Platform-specific biases in geographic coverage, represented populations, represented activities etc. (Ruths & Pfeffer, 2014; Toivonen et al., 2019).
Sports tracking data	Data from location-based applications that track personal sports activitiesFor example: Strava, SportsTracker, MapMyFitness	<ul style="list-style-type: none"> <li>- Browsing the data online (individual tracks/heatmap)</li> <li>- Programmatic access through APIs</li> <li>- Buying a pre-processed data set</li> <li>- Access varies between platforms</li> <li>- Research collaboration agreement with an operator</li> <li>- Buying a pre-processed data set</li> </ul>	Potentially global extent depending on platform popularity.	Accuracy varies according to data product and user's device: <ul style="list-style-type: none"> <li>- Raw data as GPS tracks</li> <li>- Aggregated data as road segments</li> </ul>	In raw data, timestamp reflects the actual time of the event. In aggregated data, temporal information might be limited to a specific time interval.	Data might contain classification according to activity type (e.g. biking, running) and purpose (e.g. commuting).	User ID, age, gender	<ul style="list-style-type: none"> <li>- WHERE are the most popular routes for bikers?</li> <li>- WHEN is the daily activity peak in each area?</li> </ul>	Active athletes and younger male population overrepresented (Jestico et al., 2016; Oksanen et al., 2015).
Mobile phone data	Data logged by mobile phone operators	<ul style="list-style-type: none"> <li>- Research collaboration agreement with an operator</li> <li>- Buying a pre-processed data set</li> </ul>	Extent is often limited to country borders; roaming data allow detection of international visitors.	Accuracy varies according to base station network density and data product: <ul style="list-style-type: none"> <li>- Antenna/cell-level</li> <li>- Aggregated and interpolated data</li> </ul>	In raw data, timestamp reflects the actual time of the event. In aggregated data, temporal information might be limited to a specific time interval.	No data originally available, but activities may be interpreted based on spatio-temporal analysis.	Mobile phone subscription details (if available)	<ul style="list-style-type: none"> <li>- WHERE and WHEN do people spend time in green spaces?</li> <li>- From WHERE do people arrive to green spaces?</li> </ul>	Spatial interpolation of mobile phone data can influence the outcome (Järvi et al., 2017). City centres are often overestimated by mobile phone data due to high mobile phone usage and ownership.
PPGIS data	Data collected using geospatial technologies to support public participation in public processes	<ul style="list-style-type: none"> <li>- Conducting a web map survey</li> <li>- Accessing existing data from open data portals</li> </ul>	Potentially global extent, but PPGIS surveys are often conducted from a limited study area.	Accuracy of point markers, line and polygon features varies according to question setting, available technology, zoom level and local knowledge of the user.	Timestamp reflects the time of data entry. Questionnaire can include more detailed questions about time.	Depends on the questionnaire.	Depends on the questionnaire; often contains basic background information.	<ul style="list-style-type: none"> <li>- WHAT do people value in green spaces, and WHO are they?</li> <li>- WHY are some green spaces valued more than others?</li> </ul>	Response bias (Brown, 2017); for example, more educated and enthusiastic citizens might be overrepresented among both randomly sampled and volunteer participants.



## 5.2. Exploring spatial and temporal patterns of green space use

Regarding the spatial dimension, all four data types provide relevant information about urban green spaces. Different data types highlight different hotspots and the spatial overlap between the different data sets is relatively low among the sample data sets based on the jaccard index. Spatial patterns in social media data highlight the popular and meaningful places for leisure time green space use among locals and tourists. Social media data hotspots are mostly concentrated near the city centre and popular outdoor destinations, for example, in the Helsinki Central Park, Töölö Bay, Vanhakaupunki bay and Fortress of Suomenlinna (Figs. 3 and 4). Specific hotspots depend on the platform. The spatial accuracy attached to points-of-interest likely drives the spatial pattern of Instagram and Twitter data.

Other data types are more spread out in comparison to social media data. Sports tracking data provide information on the highest spatial accuracy, revealing the use patterns across the path network based on the users' GPS tracks. Both sports tracking data and mobile phone data have the potential to capture actual routes in green spaces at different times of the day. Mobile phone data are especially fit for analysing population densities in bigger park areas such as the Central Park in Helsinki. It is, however, difficult to analyse the use patterns in smaller parks or at park edges based on mobile phone data due to the limited spatial resolution of the original network-based data. PPGIS data hotspots from the sample data sets reflect the areas from where residents wanted to share insights (Kahila-Tani et al., 2016) but did not necessarily visit. Also the temporal mismatch among some of the data sources might be reflected in the spatial patterns.

Temporal analyses of social media data, sports tracking data and mobile phone data further demonstrate the different characteristics of the data sources and the importance of analysing several data sources together, as each of them reveal different types of use patterns at distinct temporal scales. Sports tracking data and mobile phone data track people's locations continuously, providing information on the diurnal park use at a finer temporal scale, including commuters passing through green spaces. The sports tracking data is limited to physical activities, highlighting the times of active exercise. Social media data, on the other hand, show clearly the leisure time use patterns, highlighting evenings and weekends when people have time to enjoy the park and share their experiences. Despite biases, insights collected from various data sources provide detailed and dynamic views on the spatial and temporal patterns of urban green space use, or dynamics of people in cities in general (Järv et al., 2018).

## 5.3. Analysing activities in green spaces

Understanding what people do in green spaces or why they visit parks calls for more in-depth information beyond locations and time-stamps. For example, sports tracking data often include information about the type of activity but are limited to certain sport activities, such as walking, jogging and cycling. Mobile phone data do not generally contain any direct information about people's activities, although comparing location data to other geographical data sets may give hints about what activities are taking place (Ahas, Silm, et al., 2010). Textual and visual content in social media data or well-designed questions in PPGIS have been increasingly used for understanding environmental preferences (Brown & Kytä, 2014; Hausmann et al., 2018).

In the case of Helsinki, content analysis of social media data revealed a wide range of activities in green spaces. While we had no ground-truth data about the activities present in our study area, previous studies have shown that social media content likely captures the most popular activities in green spaces (Hausmann et al., 2018; Heikinheimo et al., 2017). Social media data are particularly useful for revealing emerging activities that are characteristic to a specific area, such as tree climbing in our study. However, new and special activities may become overrepresented on social media, as people are more prone

to post about them than daily activities like walking to work. The content analysis in this study was conducted manually, but automated content analysis methods are developing rapidly and provide more efficient tools for analysing large amounts of data (Richards & Tunçer, 2018; Toivonen et al., 2019). While automated content analysis of social media data helps to gain insights about revealed activities in a cost-efficient way, more traditional surveys and well-designed PPGIS studies remain relevant in recording stated activities and motivations in every day green space use.

## 5.4. Characterizing green space users

Information about the characteristics of green space users (those who have produced the data) was rather restricted in all the sample data sources. Such information may come with the data, or background characteristics can be derived based on further analysis. We sought new information about this question by looking at different language groups in social media data. Across all parks in Helsinki, Finnish was the most popular language, but the language distribution varied across green spaces. Other sample data sets had only aggregated information or no information available about the users' language preferences.

In general, there are several ways to characterize park visitors based on user-generated geographic information. Further information about the users, such as age group, gender and place of residence, could be derived from social media data through user profile information or additional analysis (Toivonen et al., 2019). Sports tracking applications and mobile phone operators possess information about the account holder. For example, mobile phone roaming data can ideally reveal the movement of international visitors. However, personal data is predominantly not accessible to researchers due to privacy regulations, although exceptions do exist (Järv et al., 2015). In contrast, PPGIS surveys enable the collection of demographic background information, such as age and gender, from the respondents, but the surveys usually reach only locals. Combining user-generated data with traditional surveys on a sample population may help to gain reliable background information about the users (Ahas, Aasa, Silm, Tiru, 2010). For example, collecting activity tracks from recruited participants allows linking the data with additional information from surveys for a focused sample of users (Vich et al., 2019).

## 5.5. Limitations with user-generated data

All user-generated data sets and related analysis come with biases that should be recognised in research (boyd et al., 2012). Three main issues require attention with the types of data presented in this study: the representativeness of, access to and ethical use of user-generated content.

None of the user-generated data sources included in this study were a random sample of green space visitors or their activities. The data were mostly a self-selected sample with inherent biases that can be difficult to measure and correct in the absence of ground-truth data. Popularity of social media platforms varies according to age group, and not everyone geotags their posts or shares content openly (Heikinheimo et al., 2017). Sports tracking data are known to be biased towards the males over women (Oksanen et al., 2015), and the applications are mostly used by sports enthusiasts having different activity patterns compared to the rest of the population. Mobile phone data from a major operator provides the best representation of the urban population in countries where the penetration rate of mobile phones is high (Ahas, Silm, Saluveer, & Järv, 2009). PPGIS surveys usually aim at a representative sample, but as they are typically short in duration and response rates may be low and biased, good representation is not easy to achieve (Brown, 2017). Popular events and other local conditions might also affect the observed patterns. For example, one user had shared hundreds of photos from a single running event in our Flickr data set. Additionally, it is important to remember that user-generated

geographic information depicts actions that took place, but does not reveal why some places were not visited or not marked on the map.

Data access is often the biggest obstacle for taking advantage of user-generated data in research (Ahas et al., 2008; Toivonen et al., 2019). Passively or indirectly contributed data sources from, for example, social media services, activity tracking apps or mobile operators are maintained by private companies. Private companies do not have a requirement to provide data access on a continuous basis, and the technological solutions for accessing the data may change or close without prior notice. For example, changes in social media APIs pose challenges for continuous data collection for research purposes (Lomborg & Bechmann, 2014). Sample data sets used in this study are not all from the same time period due to data access issues (see Supplement S1 for more details) limiting their comparability. Restricted access to proprietary data also limits the transparency and openness of the research process.

There are many ethical questions regarding the use of user-generated data sets in research which should be acknowledged (Zook, Barocas, boyd, Crawford, Keller, Gangadharan, Goodman, Hollander, Koenig, Metcalf, Narayanan, Nelson, & Pasquale, 2017). User-generated geographic information is often personal data, and analysing it without consent can be ethically problematic (Zwitter, 2014). Actively contributed data, such as PPGIS surveys have the possibility to ask for relevant consent directly from the study participants, while publicly shared social media data or mobile phone records are often used without the direct permission from the user. As the field is relatively new, legislation and ethical guidelines on using publicly but passively contributed data sets in research are still immature. Legislation also varies between countries and regions. Therefore, it is in the responsibility of researchers and the research community to take appropriate measures to safeguard sensitive information when managing and analysing personal data. This includes guarding against the re-identification of individuals from data (Zook et al., 2017), and minimizing the amount of data that is stored.

## 6. Conclusions

In this paper, we compared the ability of different user-generated data sets to provide information on where, when and how people use and value urban green spaces. Our data source comparisons suggest that social media data, sports tracking data, mobile phone data and PPGIS data can provide valuable insights for understanding urban green space use and preferences, each contributing to the understanding in their own way. While some of the data sources are better suited to answer specific questions, the optimal approach would be to incorporate insights from different types of data as each of the data sources have their limitations. Sports tracking data and mobile phone data help monitoring green space use on a fine temporal resolution throughout the day. Social media data offer a cost-effective source of information about popular and emerging leisure time activities among tourists and locals. PPGIS surveys can be designed to fill in specific information gaps. Despite the evident limitations in user-generated data sets, they are often the best available information about activities and preferences in green spaces.

## CRedit authorship contribution statement

**Vuokko Heikinheimo:** Conceptualization, Data curation, Investigation, Formal analysis, Visualization, Writing - original draft. **Henrikki Tenkanen:** Data curation, Investigation, Writing - review & editing. **Claudia Bergroth:** Data curation, Formal analysis, Writing - review & editing, Investigation. **Olle Järv:** Methodology, Writing - review & editing. **Tuomo Hiippala:** Investigation, Formal analysis, Writing - review & editing. **Tuuli Toivonen:** Conceptualization, Writing - original draft, Supervision.

## Acknowledgements

This work is a part of the project Social Media Data for Conservation Science (SoMeCon), funded by the KONE Foundation. H.T. thanks the Katumetro program for funding. T.T. thanks the Osk. Huttunen Foundation and Clare Hall, Cambridge for support. The authors would like to thank the City of Helsinki and Helsinki Region Infoshare for providing open access to the green space polygons and PPGIS data sets. We also thank the City of Helsinki for acquiring the Strava Metro data set, and Elisa Oyj for providing the mobile phone data. We are grateful to Reetta Välimäki for assisting with image classification. We would also like to thank the three anonymous reviewers for their constructive comments on an earlier version of this paper.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.landurbplan.2020.103845>.

## References

- Ahas, R., Aasa, A., Roose, A., Mark, Ü., & Silm, S. (2008). Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. *Tourism Management*, 29(3), 469–486. <https://doi.org/10.1016/j.tourman.2007.05.014>.
- Ahas, R., Aasa, A., Silm, S., & Tiru, M. (2010). Daily rhythms of suburban commuters' movements in the Tallinn metropolitan area: Case study with mobile positioning data. *Transportation Research Part C*, 18(1), 45–54. <https://doi.org/10.1016/j.trc.2009.04.011>.
- Ahas, R., Silm, S., Järv, O., Saluveer, E., & Tiru, M. (2010). Using mobile positioning data to model locations meaningful to users of mobile phones. *Journal of Urban Technology*, 17(1), 3–27. <https://doi.org/10.1080/10630731003597306>.
- Ahas, R., Silm, S., Saluveer, E., & Järv, O. (2009). Modelling home and work locations of populations using passive mobile positioning data. *Location based services and telecartography II* (pp. 301–315). Heidelberg: Springer, Berlin. [https://doi.org/10.1007/978-3-540-87393-8\\_18](https://doi.org/10.1007/978-3-540-87393-8_18).
- Anguelovski, I., Cole, H., Connolly, J., & Triguero-Mas, M. (2018). Do green neighbourhoods promote urban health justice? *The Lancet Public Health*: Vol. 3, (pp. e270–). Elsevier. [https://doi.org/10.1016/S2468-2667\(18\)30096-3](https://doi.org/10.1016/S2468-2667(18)30096-3).
- Baldwin, T., & Lui, M. (2010). Language Identification: The Long and the Short of the Matter. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 229–237).
- Batty, M. (2013). Big data, smart cities and city planning. *Dialogues in Human Geography*, 3(3), 274–279. <https://doi.org/10.1177/2043820613513390>.
- Becker, R. A., Caceres, R., Hanson, K., Loh, J. M., Urbanek, S., Varshavsky, A., & Volinsky, C. (2011). A Tale of One City: Using cellular network data for urban planning. *IEEE Pervasive Computing*, 10(4), 18–26. <https://doi.org/10.1109/MPRV.2011.44>.
- Bergroth, C. (2019). Uncovering population dynamics using mobile phone data: the case of Helsinki Metropolitan Area [Master's thesis, University of Helsinki]. <http://urn.fi/URN:NBN:fi:huib-201905272171>.
- Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5, 135–146. DOI:1511.09249v1.
- boyd, danah, & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15(5), 662–679. DOI:10.1080/1369118X.2012.678878.
- Brown, G. (2012). An empirical evaluation of the spatial accuracy of public participation GIS (PPGIS) data. *Applied Geography*, 34, 289–294. <https://doi.org/10.1016/j.apgeog.2011.12.004>.
- Brown, G. (2017). A review of sampling effects and response bias in internet participatory mapping (PPGIS/PGIS/VGI). *Transactions in GIS*, 21(1), 39–56. <https://doi.org/10.1111/tgis.12207>.
- Brown, G., & Kyttä, M. (2014). Key issues and research priorities for public participation GIS (PPGIS): A synthesis based on empirical research. *Applied Geography*, 46, 122–136. <http://www.sciencedirect.com/science/article/pii/S0143622813002531>.
- Brown, G., Schebella, M. F., & Weber, D. (2014). Using participatory GIS to measure physical activity and urban park benefits. *Landscape and Urban Planning*, 121, 34–44. <https://doi.org/10.1016/j.landurbplan.2013.09.006>.
- Burkhard, B., Kroll, F., Nedkov, S., & Müller, F. (2012). Mapping ecosystem service supply, demand and budgets. *Ecological Indicators*, 21, 17–29.
- City of Helsinki. (2018). Helsinki facts and figures 2018. <http://materialbank.myhelsinki.fi/media/2742>.
- City of Helsinki. (2019). Register of public areas in the City of Helsinki.
- Crooks, A., Pfoser, D., Jenkins, A., Croitoru, A., Stefanidis, A., Smith, D., ... Lamprianidis, G. (2015). Crowdsourcing urban form and function. *International Journal of Geographical Information Science*, 29(5), 720–741. <https://doi.org/10.1080/13658816.2014.977905>.
- Daily, G., Postel, S., Bawa, K., & Kaufman, L. (1997). *Nature's Services: Societal Dependence On Natural Ecosystems*. Bibliovault OAI Repository, the University of Chicago Press.

- DESA/UN-WUP. (2018). World Urbanization Prospects: The 2018 Revision. Department of Economic and Social Affairs, Population Division, 2. <https://population.un.org/wup/Publications/Files/WUP2018-KeyFacts.pdf>.
- Dewulf, B., Neutens, T., Lefebvre, W., Seynaeve, G., Vanpoucke, C., Beckx, C., & Van de Weghe, N. (2016). Dynamic assessment of exposure to air pollution using mobile phone data. *International Journal of Health Geographics*, 15(1), 14. <https://doi.org/10.1186/s12942-016-0042-z>.
- Di Minin, E., Tenkanen, H., & Toivonen, T. (2015). Prospects and challenges for social media data in conservation science. *Frontiers in Environmental Science*, 3. <https://doi.org/10.3389/fenvs.2015.00063>.
- Donahue, M. L., Keeler, B. L., Wood, S. A., Fisher, D. M., Hamstead, Z. A., & McPhearson, T. (2018). Using social media to understand drivers of urban park visitation in the Twin Cities, MN. *Landscape and Urban Planning*, 175, 1–10. <https://doi.org/10.1016/j.landurbplan.2018.02.006>.
- Engemann, K., Pedersen, C. B., Arge, L., Tsirogiannis, C., Mortensen, P. B., & Svenning, J.-C. (2019). Residential green space in childhood is associated with lower risk of psychiatric disorders from adolescence into adulthood. *Proceedings of the National Academy of Sciences of the United States of America*, 116(11), 5188–5193. <https://doi.org/10.1073/pnas.1807504116>.
- Fisher, D. M., Wood, S. A., Roh, Y.-H., Kim, C.-K., Fisher, D. M., Wood, S. A., ... Kim, C.-K. (2019). The geographic spread and preferences of tourists revealed by user-generated information on Jeju Island, South Korea. *Land*, 8(5), 73. <https://doi.org/10.3390/land8050073>.
- Ghermandi, A., & Sinclair, M. (2019). Passive crowdsourcing of social media in environmental research: A systematic map. *Global Environmental Change*, 55, 36–47. <https://doi.org/10.1016/j.gloenvcha.2019.02.003>.
- Gómez-Baggethun, E., Gren, Å., Barton, D. N., Langemeyer, J., McPhearson, T., O'Farrell, P., ... Kremer, P. (2013). Urban ecosystem services. *Urbanization, biodiversity and ecosystem services: Challenges and opportunities* (pp. 175–251). Netherlands: Springer. [https://doi.org/10.1007/978-94-007-7088-1\\_11](https://doi.org/10.1007/978-94-007-7088-1_11).
- Guo, S., Song, C., Pei, T., Liu, Y., Ma, T., Du, Y., ... Wang, Y. (2019). Accessibility to urban parks for elderly residents: Perspectives from mobile phone data. *Landscape and Urban Planning*, 191, 103642. <https://doi.org/10.1016/j.landurbplan.2019.103642>.
- Haaland, C., & van den Bosch, C. K. (2015). Challenges and strategies for urban green-space planning in cities undergoing densification: A review. *Urban Forestry & Urban Greening*, 14(4), 760–771. <https://doi.org/10.1016/j.ufug.2015.07.009>.
- Hamstead, Z. A., Fisher, D., Ilieva, R. T., Wood, S. A., McPhearson, T., & Kremer, P. (2018, February 2). Geolocated social media as a rapid indicator of park visitation and equitable park access. *Computers, Environment and Urban Systems*. DOI:10.1016/j.compenvurbsys.2018.01.007.
- Hausmann, A., Toivonen, T., Slotow, R., Tenkanen, H., Moilanen, A., Heikinheimo, V., & Di Minin, E. (2018). Social media data can be used to understand tourists' preferences for nature-based experiences in protected areas. *Conservation Letters*, 11(1), <https://doi.org/10.1111/conl.12343>.
- Heikinheimo, V., Di Minin, E., Tenkanen, H., Hausmann, A., Erkkonen, J., & Toivonen, T. (2017). User-generated geographic information for visitor monitoring in a National Park: A comparison of social media data and visitor survey. *ISPRS International Journal of Geo-Information*, 6(3), 85. <https://doi.org/10.3390/ijgi6030085>.
- Hieppala, T., Hausmann, A., Tenkanen, H., & Toivonen, T. (2019). Exploring the linguistic landscape of geotagged social media content in urban environments. *Digital Scholarship in the Humanities*, 34(2), 290–309. <https://doi.org/10.1093/dlsc/fqy049>.
- Hochmair, H. H., Juhász, L., & Cvetojevic, S. (2018). Data quality of points of interest in selected mapping and social media platforms. In *Progress in Location Based Services 2018* (vol. Lecture Notes in Geoinformation and Cartography) (pp. 293–313). Springer, Cham. DOI:10.1007/978-3-319-71470-7\_15.
- Huang, Q., & Wong, D. W. S. (2016). Activity patterns, socioeconomic status and urban spatial structure: What can social media data tell us? *International Journal of Geographical Information Science*, 30(9), 1873–1898. <https://doi.org/10.1080/13658816.2016.1145225>.
- Ilieva, R. T., & McPhearson, T. (2018). Social-media data for urban sustainability. *Nature Sustainability*, 1(10), 553–565. <https://doi.org/10.1038/s41893-018-0153-6>.
- Iqbal, M. S., Choudhury, C. F., Wang, P., & González, M. C. (2014). Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, 40, 63–74. <https://doi.org/10.1016/j.trc.2014.01.002>.
- Ives, C. D., Oke, C., Hehir, A., Gordon, A., Wang, Y., & Bekessy, S. A. (2017). Capturing residents' values for urban green space: Mapping, analysis and guidance for practice. *Landscape and Urban Planning*, 161, 32–43.
- Jaakkola, M., Böhlting, A., Nicklén, M., & Lämsä, A. (2016). Helsingin viher- ja virkistysverkoston kehittämissuunnitelma VISTRA osa II. Helsingin Kaupunkisuunnitteluviraston Asemakaavaosaston Selvityksiä, 2016(2).
- Jankowska, M. M., Schipperijn, J., & Kerr, J. (2015). A framework for using GPS data in physical activity and sedentary behavior studies. *Exercise and Sport Sciences Reviews*, 43(1), 48–56. <https://doi.org/10.1249/JES.0000000000000035>.
- Järvi, O., Muurisepp, K., Ahas, R., Derudder, B., & Witlox, F. (2015). Ethnic differences in activity spaces as a characteristic of segregation: A study based on mobile phone usage in Tallinn, Estonia. *Urban Studies*, 52(14), 2680–2698. <https://doi.org/10.1177/0042098014550459>.
- Järvi, O., Tenkanen, H., Salonen, M., Ahas, R., & Toivonen, T. (2018). Dynamic cities: Location-based accessibility modelling as a function of time. *Applied Geography*, 95, 101–110. <https://doi.org/10.1016/j.apgeog.2018.04.009>.
- Järvi, O., Tenkanen, H., & Toivonen, T. (2017). Enhancing spatial accuracy of mobile phone data using multi-temporal dasymetric interpolation. *International Journal of Geographical Information Science*, 31(8), 1630–1651. <https://doi.org/10.1080/13658816.2017.1287369>.
- Jenks, M., & Jones, C. A. (2009). *Dimensions of the sustainable city*. Springer Science & Business Media.
- Jestico, B., Nelson, T., & Winters, M. (2016). Mapping ridership using crowdsourced cycling data. *Journal of Transport Geography*, 52, 90–97. <https://doi.org/10.1016/j.jtrangeo.2016.03.006>.
- Kahila-Tani, M., Broberg, A., Kyttä, M., & Tyger, T. (2016). Let the citizens map—Public participation GIS as a planning support system in the Helsinki master plan process. *Planning Practice & Research*, 31(2), 195–214. <https://doi.org/10.1080/02697459.2015.1104203>.
- Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, 1(1), 1–12. <https://doi.org/10.1177/2053951714528481>.
- Korpilo, S., Virtanen, T., & Lehtäväirta, S. (2017). Smartphone GPS tracking—Inexpensive and efficient data collection on recreational movement. *Landscape and Urban Planning*, 157, 608–617. <https://doi.org/10.1016/j.landurbplan.2016.08.005>.
- Kytä, M., Broberg, A., Haybatollahi, M., & Schmidt-Thomé, K. (2016). Urban happiness: Context-sensitive study of the social sustainability of urban settings. *Environment and Planning B: Planning and Design*, 43(1), 34–57. <https://doi.org/10.1177/0265813515600121>.
- Kytä, M., Broberg, A., Tzoulas, T., & Snabb, K. (2013). Towards contextually sensitive urban densification: Location-based softGIS knowledge revealing perceived residential environmental quality. *Landscape and Urban Planning*, 113, 30–46. <https://doi.org/10.1016/j.landurbplan.2013.01.008>.
- Laatikainen, T., Tenkanen, H., Kyttä, M., & Toivonen, T. (2015). Comparing conventional and PPGIS approaches in measuring equality of access to urban aquatic environments. *Landscape and Urban Planning*, 144, 22–33. <https://doi.org/10.1016/j.landurbplan.2015.08.004>.
- Lee, A. C. K., & Maheswaran, R. (2011). The health benefits of urban green spaces: A review of the evidence. *Journal of Public Health*, 33(2), 212–222. <https://doi.org/10.1093/pubmed/fdq068>.
- Lehtomäki, J., Tuominen, S., Toivonen, T., & Leinonen, A. (2015). What data to use for forest conservation planning? A comparison of coarse open and detailed proprietary forest inventory data in Finland. *PLoS One*, 10(8), e0135926. <https://doi.org/10.1371/journal.pone.0135926>.
- Lendák, I. (2016). Mobile crowd-sensing in the Smart City. In C. Capineri, M. Haklay, H. Huang, V. Antoniou, J. Kettunen, F. Ostermann, & R. Purves (Eds.), *European Handbook of Crowdsourced Geographic Information* (p. 353).
- Levin, N., Lechner, A. M., & Brown, G. (2017). An evaluation of crowdsourced information for assessing the visitation and perceived importance of protected areas. *Applied Geography*, 79, 115–126. <https://doi.org/10.1016/j.apgeog.2016.12.009>.
- Lomborg, S., & Bechmann, A. (2014). Using APIs for data collection on social media. *The Information Society*, 30(4), 256–265. <https://doi.org/10.1080/01972243.2014.915276>.
- McCay-Peet, L., & Quan-Haase, A. (2017). What is social media and what questions can social media research help us answer. In L. Sloan & A. Quan-Haase (Eds.), *The SAGE handbook of social media research methods* (pp. 13–26). SAGE.
- MEA. (2005). MEA = Millennium Ecosystem Assessment. Ecosystems and human well-being (Vol. 5). Island Press Washington, DC.
- Mytton, O. T., Townsend, N., Rutter, H., & Foster, C. (2012). Green space and physical activity: An observational study using Health Survey for England data. *Health & Place*, 18(5), 1034–1041.
- Ngom, R., Gosselin, P., & Blais, C. (2016). Reduction of disparities in access to green spaces: Their geographic insertion and recreational functions matter. *Applied Geography*, 66, 35–51. <https://doi.org/10.1016/j.apgeog.2015.11.008>.
- Norman, P., & Pickering, C. M. (2017). Using volunteered geographic information to assess park visitation: Comparing three on-line platforms. *Applied Geography*, 89, 163–172. <https://doi.org/10.1016/j.apgeog.2017.11.001>.
- Oksanen, J., Bergman, C., Sainio, J., & Westerholm, J. (2015). Methods for deriving and calibrating privacy-preserving heat maps from mobile sports tracking application data. *Journal of Transport Geography*, 48, 135–144. <https://doi.org/10.1016/j.jtrangeo.2015.09.001>.
- Pietrzyk-Kaszyńska, A., Czepkiewicz, M., & Kronenberg, J. (2017). Eliciting non-monetary values of formal and informal urban green spaces using public participation GIS. *Landscape and Urban Planning*, 160, 85–95. <https://doi.org/10.1016/j.landurbplan.2016.12.012>.
- Richards, D. R., & Tunçer, B. (2018, September 18). Using image recognition to automate assessment of cultural ecosystem services from social media photographs. *Ecosystem Services*. <https://doi.org/10.1016/j.ecoser.2017.09.004>.
- Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, 346(6213), 1063–1064. <https://doi.org/10.1126/science.346.6213.1063>.
- See, L., Mooney, P., Foody, G., Bastin, L., Comber, A., Estima, J., ... Rutzinger, M. (2016). Crowdsourcing, Citizen Science or Volunteered Geographic Information? The Current State of Crowdsourced Geographic Information. *ISPRS International Journal of Geo-Information*, 5(5), 55. <https://doi.org/10.3390/ijgi5050055>.
- Shelton, T., Poorthuis, A., & Zook, M. (2015). Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information. *Landscape and Urban Planning*, 142, 198–211. <https://doi.org/10.1016/j.landurbplan.2015.02.020>.
- Statistics Finland. (2019). Official Statistics of Finland (OSF): Population structure [e-publication]. ISSN = 1797-5395. [http://www.stat.fi/til/vaerak/index\\_en.html](http://www.stat.fi/til/vaerak/index_en.html).
- Steiger, E., Westerholt, R., Resch, B., & Zipf, A. (2015). Twitter as an indicator for whereabouts of people? Correlating Twitter with UK census data. *Computers, Environment and Urban Systems*, 54, 255–265. <https://doi.org/10.1016/j.compenvurbsys.2015.09.007>.
- Tarnanen, A., Salonen, M., Willberg, E., & Toivonen, T. (2017). Pyöräilyn reitit ja sujuvuus. *Kaupunkiympäristön julkaisuja/Kaupunkiympäristön julkaisuja*, 2017(16), 122.
- Tenkanen, H., Di Minin, E., Heikinheimo, V., Hausmann, A., Herbst, M., Kajala, L., & Toivonen, T. (2017). Instagram, Flickr, or Twitter: Assessing the usability of social media data for visitor monitoring in protected areas. *Scientific Reports*, 7(1), 17615.

- <https://doi.org/10.1038/s41598-017-18007-4>.
- Toivonen, T., Heikinheimo, V., Fink, C., Hausmann, A., Hiippala, T., Järv, O., ... Di Minin, E. (2019). Social media data for conservation science: A methodological overview. *Biological Conservation*, 233, 298–315. <https://doi.org/10.1016/J.BIOCON.2019.01.023>.
- Tomita, A., Vandormael, A. M., Cuadros, D., Di Minin, E., Heikinheimo, V., Tanser, F., ... Burns, J. K. (2017). Green environment and incident depression in South Africa: A geospatial analysis and mental health implications in a resource-limited setting. *The Lancet Planetary Health*, 1(4), e152–e162. [https://doi.org/10.1016/S2542-5196\(17\)30063-3](https://doi.org/10.1016/S2542-5196(17)30063-3).
- Tyrväinen, L., Mäkinen, K., & Schipperijn, J. (2007). Tools for mapping social values of urban woodlands and other green areas. *Landscape and Urban Planning*, 79(1), 5–19. <https://doi.org/10.1016/j.landurbplan.2006.03.003>.
- Vich, G., Marquet, O., & Miralles-Guasch, C. (2019). Green streetscape and walking: Exploring active mobility patterns in dense and compact cities. *Journal of Transport & Health*, 12, 50–59. <https://doi.org/10.1016/J.JTH.2018.11.003>.
- Wang, Y., Kotze, D. J., Vierikko, K., & Niemelä, J. (2019). What makes urban greenspace unique – Relationships between citizens' perceptions on unique urban nature, biodiversity and environmental factors. *Urban Forestry and Urban Greening*, 42, 1–9. <https://doi.org/10.1016/j.ufug.2019.04.005>.
- Ward Thompson, C., Roe, J., Aspinall, P., Mitchell, R., Clow, A., & Miller, D. (2012). More green space is linked to less stress in deprived communities: Evidence from salivary cortisol patterns. *Landscape and Urban Planning*, 105(3), 221–229. <http://linkinghub.elsevier.com/retrieve/pii/S0169204611003665>.
- Wolch, J. R., Byrne, J., & Newell, J. P. (2014). Urban green space, public health, and environmental justice: The challenge of making cities “just green enough”. *Landscape and Urban Planning*, 125, 234–244. <https://doi.org/10.1016/j.landurbplan.2014.01.017>.
- Wood, S. A., Guerry, A. D., Silver, J. M., & Lacayo, M. (2013). Using social media to quantify nature-based tourism and recreation. *Scientific Reports*, 3.
- Zhou, X., & Wang, Y.-C. (2011). Spatial-temporal dynamics of urban green space in response to rapid urbanization and greening policies. *Landscape and Urban Planning*, 100(3), 268–277. <https://doi.org/10.1016/J.LANDURBPLAN.2010.12.013>.
- Zook, M., Barocas, S., Boyd, D., Crawford, K., Keller, E., Gangadharan, S. P., Goodman, A., Hollander, R., Koenig, B. A., Metcalf, J., Narayanan, A., Nelson, A., & Pasquale, F. (2017). Ten simple rules for responsible big data research. In F. Lewitter (Ed.), *PLoS Computational Biology* (Vol. 13, Issue 3, p. e1005399). Springer. DOI:10.1371/journal.pcbi.1005399.
- Zwitter, A. (2014). Big Data ethics. *Big Data & Society*, 1(2), 205395171455925. DOI:10.1177/2053951714559253.